

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization International Bureau



(43) International Publication Date
3 February 2005 (03.02.2005)

PCT

(10) International Publication Number
WO 2005/010642 A2

(51) International Patent Classification⁷:

G06F

(74) Agent: G.E. EHRLICH (1995) LTD., 11 Menachem Begin Street, 52 521 Ramat Gan (IL).

(21) International Application Number:

PCT/IL2004/000704

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(22) International Filing Date: 1 August 2004 (01.08.2004)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

60/491,235 31 July 2003 (31.07.2003) US
60/510,553 14 October 2003 (14.10.2003) US

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(71) Applicants (for all designated States except US): RAMOT AT TEL AVIV UNIVERSITY LTD. [IL/IL]; P.O. Box 39296, 61 392 Tel Aviv (IL). CORNELL RESEARCH FOUNDATION, INC. [US/US]; Cornell Business & Technology Park, 20 Thornwood Drive, Suite 105, Ithaca, New York 14850 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): EDELMAN, Shimon [IL/US]; 106 Brandywine Drive, Ithaca, New York 14850-1708 (US). HORN, David [IL/IL]; 28 Gutman Nachum Street, 69343 Tel Aviv (IL). RUPPIN, Eytan [IL/IL]; 44 Lilach Street, 71 908 Reut (IL). SOLAN, Tsach [IL/IL]; 10 Lea Street, 69 412 Tel Aviv (IL).

Published:

— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2005/010642 A2

(54) Title: METHOD AND APPARATUS FOR LEARNING, RECOGNIZING AND GENERALIZING SEQUENCES

(57) Abstract: A method of generalizing a dataset having a plurality of sequences defined over a lexicon of tokens is provided. The method comprises: searching over the dataset for similarity sets, where each similarity set comprises a plurality of segments of size L having L-S common tokens and S uncommon tokens; and defining a plurality of equivalence classes corresponding to uncommon tokens of at least one similarity set. The method may further comprise a step in which a plurality of significant patterns are extracted, where each significant pattern corresponds to a most significant partial overlap between one sequence of the dataset and other sequences of the dataset. In one embodiment, a generalized dataset represented by a graph or a forest is constructed, and can be realized as a context-free grammar. The graph or forest can be used for generating sequences and/or testing grammatical structures.